# On the Similarity of Sets of Permutations and its Applications to Genome Comparison

Anne Bergeron[1] and Jens Stoye[2]

[1] LaCIM, Université du Québec à Montréal, Canada,
anne@lacim.uqam.ca
[2] Technische Fakultät, Universität Bielefeld, Germany,
stoye@techfak.uni-bielefeld.de

**Abstract.** The comparison of genomes with the same gene content relies on our ability to compare permutations, either by measuring how much they differ, or by measuring how much they are alike. With the notable exception of the breakpoint distance, which is based on the concept of conserved adjacencies, measures of distance do not generalize easily to sets of more than two permutations. In this paper, we present a basic unifying notion, *conserved intervals*, as a powerful generalization of adjacencies, and as a key feature of genome rearrangement theories. We also show that sets of conserved intervals have elegant nesting and chaining properties that allow the development of compact graphic representations, and linear time algorithms to manipulate them.

## 1 Introduction

Gene order analysis in a set of organisms is a powerful technique for phylogenetic inference. Current methods are based on notions of *distances* between genomes, which are usually defined as the minimum number of such and such operations needed to transform one genome into the other one. Distance matrices can either be used directly as data for phylogenetic reconstruction, or in more qualitative attempts to reconstruct ancestral genomes [9]. All these methods, with the notable exception of the *breakpoint distance* [6], are closely tied to initial choices of allowable rearrangement operations. They are also *pure* distances, in the sense that similarities between genomes are purposefully ignored.

The breakpoint distance is based on the notion of conserved adjacencies. Compared to other distances, it is easy to compute, but it often fails to capture more global relations between genomes [17]. Nevertheless, conserved adjacencies have two highly desirable properties:

1. They can be defined on a set of more than two genomes, allowing for the identification of similar features in a family of organisms.
2. They are invariant under optimal rearrangement scenarios, in the sense that it is not necessary to break adjacencies to explain how a genome evolved from another one [10, 15, 21].

| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Fruit Fly | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 |
| Mosquito | 1 | 2 | 3 | 4 | 5 | 6 | 8 | 7 | 9 | −10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 |
| Silkworm | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 14 | 13 | 15 | 16 | 17 |
| Locust | 1 | 2 | 3 | 5 | 4 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 |
| Tick | 1 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | −2 | 12 | 13 | 14 | 15 | 16 | 17 |
| Centipede | 1 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | −2 | 12 | 16 | 13 | 14 | 15 | 17 |

**Table 1.** Condensed mitochondrial genomes of six Arthropoda

A first generalization of adjacencies is the notion of *common intervals* that identify subsets of genes that appear consecutively in two or more genomes [13, 22]. Common intervals identify more global relations between genomes, but often lose the invariant property of adjacencies with respect to optimal rearrangement scenarios. For example, all optimal sortings by reversals of the permutation $(1\ \ 3\ \ 2\ \ 5\ \ -4\ \ 6)$ break, in some of the intermediate permutations, the common interval $(2\ \ 3)$.

Are adjacencies the only structures that are invariant under biologically meaningful rearrangement operations? No. There exists a class of common intervals, called *conserved intervals*, that may be the best of two worlds. We will show that these structures capture both local and global properties of genomes; are invariant under most rearrangement scenarios; and their number and nature can be computed in linear time.

## 2 Permutations, Gene Order, and Rearrangements

In the following we will take for granted the simplifying hypothesis that the genes of an organism are ordered and oriented along linear or circular DNA molecules. For example the 37 mitochondrial genes of the Fruit Fly are listed in [7], with minus signs to reflect orientation, as:

```
cox1, L2, cox2, K, D, atp8, atp6, cox3, G, nad3, A, R, N, S1, E, -F,
-nad5, -H, -nad4, -nad4L, T, -P, nad6, cob, S2, -nad1, -L1, -rrnL, -V,
-rrnS, UNK, I, -Q, M, nad2, W, -C, -Y
```

The first gene is arbitrary, since mitochondrial genomes are circular molecules. When organisms with the same gene content are compared, one of them is chosen as a base organism, and all identical strips of genes are converted to integers. By extension, these are also called "genes". Table 1 presents the result of this transformation applied to the mitochondrial genomes of six *Arthropoda*, with Fruit Fly as base organism. The original 37 genes have been divided in 17 blocks: some represent isolated genes, and others represent longer strips. For example, 10 stands for `S1`, and 11 for `E, -F, -nad5, -H, -nad4, -nad4L, T, -P, nad6, cob, S2, -nad1`.

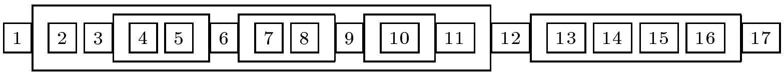
Various techniques are then used to compare the resulting permutations. The *distance* approaches focus on the differences between two particular genomes. For example, Fruit Fly differs from Mosquito by the *reversal* of gene 10, and the *transposition* of genes 7 and 8. One can count the minimal number of reversals and/or transpositions necessary to transform each genome into any other, yielding a distance matrix for the set of species. Explicit rearrangement scenarios, that is, sequences of operations that transform optimally one genome into another, are also used to reconstruct ancestral genomes.

Another approach, the *breakpoint distance*, counts the lost adjacencies between genomes. It does not rely on particular rearrangement operations or an evolutionary model, and it has an associated measure of similarity: the number of *conserved* adjacencies. For example, given the circularity of the genomes, Fruit Fly and Mosquito have 12 conserved adjacencies, and their breakpoint distance is 5.

Such a similarity measure extends easily to sets of species. For example, the first four species of Table 1 share 6 adjacencies: $[1, 2]$, $[2, 3]$, $[11, 12]$, $[15, 16]$, $[16, 17]$, and $[17, 1]$. When comparing all six species, the only left adjacency is $[17, 1]$: this lack of conserved adjacencies is a direct consequence of how the data was transformed. Does this mean that losing common adjacencies amounts to losing all common structures?

A quick glance at Table 1 reveals that the six permutations are very "similar". For example, the genes in the interval $[1, 12]$ are all the same, with small variations in their ordering. This is also true for the genes in the intervals $[3, 6]$, $[6, 9]$, $[9, 11]$, and $[12, 17]$. It turns out that such intervals, together with conserved adjacencies, play a fundamental role in rearrangement and distance theories, ancestral genome reconstructions, and phylogeny.

The following *family portrait* gives a representation of the conserved intervals of the permutations of Table 1:



This representation boxes the elements in rectangles, which can be glued together to form larger objects. It takes its roots in $PQ$-trees [8] that are used to represent sets of permutations. All permutations of Table 1 fit the representation with the following conventions: (1) free objects within a rectangle can be reordered, or can change sign, (2) connections between rectangles are fixed. This representation also captures the features that should be invariant in biologically plausible rearrangement scenarios within the family.

In order to illustrate this last point, consider the two following rearrangement scenarios that transform Silkworm into Locust using a minimal number of *reversals* (operations that reverse the elements of a consecutive block while changing their signs).

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 14 | 13 | 15 | 16 | 17 |
| 1 | 2 | 3 | -4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 14 | 13 | 15 | 16 | 17 |
| 1 | 2 | 3 | -4 | -5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 14 | 13 | 15 | 16 | 17 |
| 1 | 2 | 3 | 5 | 4 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 14 | 13 | 15 | 16 | 17 |
| 1 | 2 | 3 | 5 | 4 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | -14 | 13 | 15 | 16 | 17 |
| 1 | 2 | 3 | 5 | 4 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | -14 | -13 | 15 | 16 | 17 |
| 1 | 2 | 3 | 5 | 4 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 |

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 14 | 13 | 15 | 16 | 17 |
| 1 | 2 | 3 | 4 | -14 | -12 | -11 | -10 | -9 | -8 | -7 | -6 | -5 | 13 | 15 | 16 | 17 |
| 1 | 2 | 3 | 4 | -14 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 15 | 16 | 17 |
| 1 | 2 | 3 | 4 | -13 | -12 | -11 | -10 | -9 | -8 | -7 | -6 | -5 | 14 | 15 | 16 | 17 |
| 1 | 2 | 3 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | -4 | 14 | 15 | 16 | 17 |
| 1 | 2 | 3 | 5 | 4 | -13 | -12 | -11 | -10 | -9 | -8 | -7 | -6 | 14 | 15 | 16 | 17 |
| 1 | 2 | 3 | 5 | 4 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 |

Those two scenarios are fundamentally different, even if they both use six reversals. The right one uses much longer reversals than the left one, and the right one *breaks* conserved intervals between Silkworm and Locust in intermediate permutations, namely $[3, 6]$, $[1, 12]$, and $[12, 17]$. If a rearrangement scenario is expected to reflect the various intermediate species between Silkworm and Locust, the right one looks highly suspicious. Recent papers address these problems in various ways, for example by assigning weights to operations [1], or with probabilistic studies of the possible scenarios [16].

The two main flaws of the second scenario – long reversals and breaking conserved intervals – are closely tied: breaking conserved intervals, as we will show in Sect. 6, often involves long range operations that radically disturb a genome. In this sense, conserved intervals can be used as an intrinsic measure that allows to screen out rearrangement scenarios, or phylogenetic hypotheses, without the need of arbitrary weights or probability measures.

## 3 Conserved Intervals

This section presents a formalization of the notion of conserved intervals, together with properties that allow the development of linear time algorithms to manipulate them.

**Definition 1.** *Let $G$ be a set of signed permutations on $n$ elements. An interval $[a, b]$ is a* conserved interval *of the set $G$ if:*

*1) either $a$ precedes $b$, or $-b$ precedes $-a$, in each permutation, and*

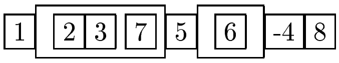
*2) the set of unsigned elements that appear between $a$ and $b$ is the same for all permutations in $G$.*

An elementary consequence of this definition is the fact that if $[a, b]$ is a conserved interval, so is $[-b, -a]$. We will consider these intervals as equivalent.

Table 1 contains several examples of conserved intervals. Their description is eased by the fact that the identity permutation belongs to the set $G$. When this is the case, all conserved intervals can be identified with their positive endpoints $a < b$, and the set of elements that appear between $a$ and $b$ is $\{a + 1, \ldots, b - 1\}$. The following example illustrates a more general case. Consider the two permutations:

$$\begin{array}{rcrrrrrrrr} P & = & 1 & 2 & 3 & 7 & 5 & 6 & -4 & 8 \\ Q & = & 1 & 7 & -3 & -2 & 5 & -6 & -4 & 8 \end{array}$$

In this example, $[1, 5]$ and $[2, 3]$ are conserved intervals, but not $[1, 6]$. The other conserved intervals of $P$ and $Q$ are $[1, -4]$, $[1, 8]$, $[5, -4]$, $[5, 8]$, and $[-4, 8]$. The diagram representation of these intervals, with respect to the permutation $P$, is:



When the identity permutation is not in $G$, it is always possible to *rename* the elements of $G$ such that conserved intervals will be intervals of consecutive elements. For example, if one composes[3] the permutations $P$ and $Q$ of the above

[3] Here, composition is understood as the standard composition of functions. Dealing with signed permutations requires the additional axiom that $P(-a) = -P(a)$.

example with the inverse permutation $P^{-1}$, the first permutation becomes the identity permutation $Id = P^{-1} \circ P$. In general, it is elementary to transform a set of conserved intervals to its equivalent up to renaming. It is a consequence of the following proposition:

**Proposition 1.** *Let $R$ be a permutation and $G$ a set of permutations, denote by $R \circ G$ the set of permutations obtained by composing each permutation in $G$ with $R$. The interval $[a, b]$ is conserved in $G$ if and only if the interval $[R(a), R(b)]$ is conserved in $R \circ G$.*

Some intervals, such as $[1, -4]$ for the set $\{P, Q\}$ in the above example, are the union of smaller intervals: $[1, -4] = [1, 5] \cup [5, -4]$. Intervals that are not unions are specially useful:

**Definition 2.** *Conserved intervals that are not the union of shorter conserved intervals are called* irreducible.

Sets of conserved intervals can be simply characterized by the corresponding set of irreducible intervals. Indeed, disjoint irreducible intervals, as highlighted in the diagram representation, are either *chained* or *nested.* The following proposition captures the basic properties of these structures.

**Proposition 2 ([5]).** *Two different irreducible conserved intervals $[a, b]$ and $[c, d]$ of a set $G$ of permutations, are either:*
*1) disjoint,*
*2) nested with different endpoints, or*
*3) overlapping on one element.*

Overlapping irreducible intervals form chains linked by their successive common elements. A chain of $k - 1$ intervals $[a_1, a_2][a_2, a_3] \ldots [a_{k-1}, a_k]$ will be denoted simply by its $k$ links $[a_1, a_2, a_3, \ldots, a_k]$. For example, $[1, 5, -4, 8]$ is a chain of the set of conserved intervals of $P$ and $Q$. A *maximal* chain is a chain that cannot be extended. We have:

**Proposition 3.** *Every irreducible conserved interval belongs to a unique maximal chain.*

One consequence of Proposition 3 is that maximal chains, as sets of links, together with isolated genes, form a partition of the set of genes. This will reveal useful to construct data structures to keep track of conserved intervals.

A set of permutations on $n$ elements can have as many as $n(n-1)/2$ conserved intervals, but at most $n-1$ irreducible intervals. These bounds are achieved with sets containing only one permutation. A key observation, that will eventually lead to linear time algorithms to compute the number of conserved intervals, is the following:

**Proposition 4.** *Each maximal chain of $k$ links contributes $k(k-1)/2$ to the total number of conserved intervals.*

Finally, we will want to construct sets of conserved intervals for the union of two sets of permutations. Definition 1 implies that the set of conserved intervals of a union of two sets of permutations is the intersection of their sets of conserved intervals. The following proposition, shown in [5], relates these sets to their respective irreducible intervals when both sets of permutations have at least one permutation in common.

**Proposition 5.** *Let $P$ be a permutation that is contained in both sets of permutations $G_1$ and $G_2$. The interval $[a,b]$ is a conserved interval of $G = G_1 \cup G_2$ if and only if there exist two chains of irreducible conserved intervals, with respect to $P$, with $k \geq 0$, $m \geq 0$:*

$$\begin{array}{ll} [a, x_1, \ldots, x_k, b] & \text{in } G_1, \\ [a, y_1, \ldots, y_m, b] & \text{in } G_2. \end{array}$$

*The interval $[a,b]$ is irreducible if and only if $\{x_1, \ldots, x_k\}$ and $\{y_1, \ldots, y_m\}$ are disjoint.*

**Variable Geometry Genomes.** Although the definition of conserved intervals was given for permutations that model genomes composed of single linear chromosomes, they can be adapted to other types of genomes. For details, see [5].

## 4 Algorithms

This section discusses three algorithms. The first one is an adaptation of an existing algorithm that computes the conserved intervals of two permutations. The second one computes the conserved intervals of a set of permutations. The third one, finally, computes the conserved intervals of two sets of permutations, directly from their two individual sets of conserved intervals.

**Conserved Intervals of Two Permutations.** Conserved intervals between two permutations are strongly related to the notion of connected components of the *overlap graph* of a signed permutation. This graph plays a fundamental role in the sorting by reversals problem [11], and the sorting by reversals and translocations problem [12]. In the last few years, linear algorithms to identify these components have been devised [2]. The following algorithm is adapted from [4], and identifies all irreducible conserved intervals[4] $[a,b]$ of a permutation $\pi$ with the identity permutation such that both $a$ and $b$ have positive sign in $\pi$. The case of negative endpoints is treated by reversing $\pi$.

For example, for the permutation

$$P = \;\; 0 \; -4 \; -3 \; -2 \;\; 5 \;\; 8 \;\; 6 \;\; 7 \;\; 9 \; -1 \;\; 10,$$

Algorithm 1 identifies the positive irreducible conserved intervals $[6,7]$, $[5,9]$, and $[0,10]$. It will identify $[2,3]$ and $[3,4]$ on the reversed permutation.

[4] In the original paper, these were called *framed common intervals.*

**Algorithm 1** (Positive irreducible intervals with the identity permutation)

```
stack 0 on $\mathcal{S}$
stack $n$ on $\mathcal{M}$
$M_0 \leftarrow n$
for $i = 1, \ldots, n$ do
  // Computation of $M_i$
  unstack from $\mathcal{M}$ all elements $m$ smaller than $|\pi_i|$
  $M_i \leftarrow m$
  stack the element $|\pi_i|$ on $\mathcal{M}$
  // Identification of irreducible intervals
  unstack from $\mathcal{S}$ all indices $s$ such that $(|\pi_i| < \pi_s$ or $|\pi_i| > M_s)$
  if $i - s = \pi_i - \pi_s$ and $M_i = M_s$ then
    output $[\pi_s, \pi_i]$
  end if
  if $\pi_i$ is positive then
    stack the index $i$ on $\mathcal{S}$
  end if
end for
```

The algorithm assumes that the input permutation is in the form $\pi = (0, \pi_1, \ldots, \pi_{n-1}, n)$. Define $M_i$ to be the nearest unsigned element of the permutation that precedes $\pi_i$ and is greater than $|\pi_i|$. (Set $M_i$ to $n$, if such an element does not exist). The following lemma relates the values of $M_i$ to conserved intervals.

**Lemma 1.** *If $[\pi_s, \pi_e]$ is a positive conserved interval of $\pi$ and the identity permutation, then $M_s = M_e$.*

The algorithm uses two stacks: $\mathcal{S}$ contains the possible start positions of conserved intervals; $\mathcal{M}$ contains possible candidates for $M_i$. The top of $\mathcal{S}$ is always denoted by $s$. The top of $\mathcal{M}$ is always denoted by $m$.

**Proposition 6 ([4, 5]).** *Algorithm 1 outputs the positive irreducible conserved intervals of a permutation $\pi$ with the identity permutation in $O(n)$ time.*

**Corollary 1.** *By applying Algorithm 1 both to $\pi = P^{-1} \circ Q$ and to the reverse of $\pi$, the irreducible conserved intervals of two permutations $P$ and $Q$ can be found in $O(n)$ time.*

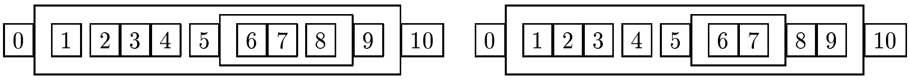
**Conserved Intervals of a Set of Permutations.** In order to find the irreducible conserved intervals of a set of permutations, the first step is to compute the irreducible intervals of each permutation with one particular permutation from the set, say $\pi_1$, using Algorithm 1, and then merge together the resulting sets of irreducible intervals. For example, computing the irreducible intervals of the set:

$$\begin{array}{rcrrrrrrrrrrr} Id & = & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ P & = & 0 & -4 & -3 & -2 & 5 & 8 & 6 & 7 & 9 & -1 & 10 \\ Q & = & 0 & 5 & -7 & -6 & 8 & 9 & 1 & 2 & 3 & -4 & 10 \end{array}$$

**Algorithm 2** (Irreducible intervals of $G_1 \cup G_2$, both containing the identity permutation)

```
stack 0 on $\mathcal{S}$
for $i = 1, \ldots, n$ do
    if there is an interval $[x, i]$ in $I_1$ then
        unstack from $\mathcal{S}$ all elements larger than $x$
    end if
    if there is an interval $[x, i]$ in $I_2$ then
        unstack from $\mathcal{S}$ all elements larger than $x$
    end if
    if $s$ and $i$ belong to the same chain both in $I_1$ and $I_2$ then
        unstack $s$ from $\mathcal{S}$ and output $[s, i]$
    end if
    if there is an interval that starts at $i$ in $I_1$, and one in $I_2$ then
        stack $i$ on $\mathcal{S}$
    end if
end for
```

would first yield the two sets of maximal chains $\{[0, 10], [2, 3, 4][5, 9], [6, 7]\}$ (of $P$ and the identity) and $\{[0, 10], [1, 2, 3], [5, 8, 9], [6, 7]\}$ (of $Q$ and the identity), respectively, in graphic representation:



Assume that each set of irreducible conserved intervals is given by its maximal chains. Since these form partitions of the genes that are endpoints of conserved intervals, there exists a data structure with the following properties: (1) For each index from 1 to $n$, it is possible to determine in constant time the interval, if any, that starts and/or ends at this index. (2) It is possible to determine in constant time if two intervals belong to the same chain.

Let $I_1$ and $I_2$ be two sets of irreducible conserved intervals of sets of permutations $G_1$ and $G_2$ that have one permutation $\pi_1$ in common. For the moment we will assume that $\pi_1$ is the identity permutation. Then Algorithm 2 finds all irreducible conserved intervals of $G_1 \cup G_2$. It uses a stack $\mathcal{S}$ that contains possible start positions – or, equivalently, elements of the identity permutation. The top of the stack $\mathcal{S}$ is always denoted by $s$.

The correctness and time complexity of Algorithm 2 are established by the following theorem, whose proof can be found in [5].

**Theorem 1.** *Algorithm 2 outputs the irreducible intervals of $G = G_1 \cup G_2$ in $O(n)$ time, given $I_1$ and $I_2$, the irreducible intervals of two sets of permutations $G_1$ and $G_2$ that both contain the identity permutation.*

**Corollary 2.** *Let $I_1$ and $I_2$ be the irreducible intervals of two sets of permutations $G_1$ and $G_2$ that both contain a permutation $P$. The irreducible intervals of $G = G_1 \cup G_2$ can be found in $O(n)$ time by applying Algorithm 2 to $I'_1 = \{[P^{-1}(a), P^{-1}(b)] \mid [a, b] \in I_1\}$ and $I'_2 = \{[P^{-1}(a), P^{-1}(b)] \mid [a, b] \in I_2\}$.*

**Corollary 3.** *The set of irreducible conserved intervals of a set of permutations $G$ can be computed in $O(|G|n)$ time and $O(n)$ additional space.*

**Conserved Intervals of Disjoint Sets.** Finally we are interested in computing the conserved intervals of two sets of permutations $G_1 = \{P_1, \ldots, P_k\}$ and $G_2 = \{Q_1, \ldots, Q_m\}$ that not necessarily have a permutation in common, given their sets of irreducible conserved intervals $I_1$ and $I_2$, respectively.

This can be done in linear time by properly combining Algorithms 1 and 2. The idea is to select one permutation from each set, say $P_1$ from $G_1$ and $Q_1$ from $G_2$, and compute the conserved intervals of these two by Algorithm 1. Then observe that the two sets $\{P_1, Q_1\}$ and $G_1 = \{P_1, \ldots, P_k\}$ have a joint permutation $P_1$, and hence their common irreducible intervals can be computed by Algorithm 2. Similarly, $\{Q_1, P_1, \ldots, P_k\}$ and $G_2 = \{Q_1, \ldots, Q_m\}$ contain a joint permutation $Q_1$, so their common irreducible intervals can also be computed by Algorithm 2.

## 5 Similarity and Distance

The number of conserved intervals of a set of permutations is a measure of similarity, but it can easily be transformed into a distance between two permutations, or two sets of permutations. The basic idea is that two sets of conserved intervals can be compared with the cardinality of their symmetric difference.

**Definition 3.** *Let $G_1$ and $G_2$ be two sets of permutations on $n$ elements, with respectively $N_1$ and $N_2$ conserved intervals. Let $N$ be the number of conserved intervals in $G_1 \cup G_2$. The* interval distance *between $G_1$ and $G_2$ is defined by $d(G_1, G_2) = N_1 + N_2 - 2N$.*

*Note:* The interval distance satisfies the fundamental properties of a mathematical distance since one can prove that the relation is symmetric, reflexive, and satisfies the *triangle inequality*: $d(G_1, G') + d(G', G_2) \geq d(G_1, G_2)$.

A detailed comparison of the interval distance with other rearrangement distances can be found in [5]. The behavior of the interval distance is a consequence of the fact that it is affected be the length – or number of genes – involved in a rearrangement operation: short reversals, for example, are less disturbing than long ones. In particular, the amount of disruption due to a single rearrangement operation can readily be computed. For example, we have the following:

**Proposition 7.** *Suppose that $P$ and $Q$ have $n$ elements, then:*

*1) if $P$ is obtained from $Q$ by reversing $k$ elements, then the interval distance between $P$ and $Q$ is $k(n-k)$;*

*2) if $P$ is obtained from $Q$ by transposing two consecutive blocks of $a$ and $b$ elements, then the interval distance between $P$ and $Q$ is $(a+b)(n-(a+b))+ab$.*

Since the interval distance is affected by length, the practice of collapsing identical strips of genes should be questioned. Indeed, as we saw in the example of Sect. 2, the integers resulting from such a transformation stand for strips of genes that vary greatly in length. We believe that whole genome comparison should use all available information, and that length of segments is relevant to the study of rearrangement scenarios, as advocated in [19].

## 6 Links With Rearrangement Theories

In Sect. 2, we gave an example of how conserved intervals could be used to evaluate optimal reversal scenarios between two genomes. Reversals are one of the many operations that are currently used to model genome evolution: the main other ones – among those that do not need to model duplication of genes – are transpositions, reverse transpositions, translocations, fusions, and fissions.

In this section, we want to characterize the rearrangement operations, or scenarios, that *preserve* conserved intervals:

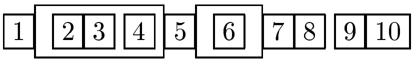
**Definition 4.** *Let $P$ and $Q$ be two permutations, and $\rho$ a rearrangement operation applied to $P$ yielding $P'$. We say that $\rho$* preserves *the conserved intervals of $P$ and $Q$ if the conserved intervals of $\{P, Q\}$ are contained in those of $\{P', Q\}$.*

Keeping in mind the graphical representation of the conserved intervals, it is easy to identify the operations that preserve conserved intervals: only rearrangements within blocks are preserving. To be more formal, note that all operations, except fusions, destroy some adjacencies that existed in the original permutation: the number and nature of these adjacencies is a key concept.

**Definition 5.** *Let $\rho$ be a rearrangement operation that transforms $P$ into $P'$. A* breakpoint *of $\rho$ is a pair of elements that are adjacent in $P$ but not in $P'$.*

In other words, breakpoints are where one has to cut $P$ in order to apply $\rho$. Reversals and translocations have 2 breakpoints, transpositions have 3, and fissions have 1.

Consider the irreducible intervals of $P$ and $P'$ with respect to $P$. Adjacencies in $P$ either belong to a (smallest) irreducible interval, or are *free.* For example, in the diagram



the adjacency $(3, 4)$ belongs to the interval $[1, 5]$, $(2, 3)$ belongs to $[2, 3]$, and $(8, 9)$ is free. Note that when two or more adjacencies belong to the same irreducible interval, then none of these adjacencies is conserved between $P$ and $P'$.

**Theorem 2 ([5]).** *Reversals, transpositions, and reverse transpositions are preserving if and only if all their breakpoints belong to the same irreducible interval, or are free. Translocations and fissions are preserving if and only if all their breakpoints are free.*

It turns out that most rearrangement operations used in optimal scenarios are indeed preserving. It is outside the scope of this paper to discuss these results in detail: they involve the *cycle* structure of a permutation, which are special subsets of the breakpoints of a permutation $P$ with respect to a permutation $P'$. The following result has been proved in various disguises in recent years [4, 11, 14]:

**Theorem 3.** *All the breakpoints of a cycle belong to the same irreducible interval.*

In the sorting by reversals theory, a *sorting* reversal is defined as a reversal that decreases the reversal distance by 1. It is shown [11, 20] that the breakpoints of sorting reversals, except for one type called *component merging*, belong to a single cycle, thus we have:

**Corollary 4.** *All sorting reversals, except component merging, are preserving.*

Component mergings are a rare type of reversals in optimal scenarios: they break at least two irreducible intervals, thus they often involve long reversals.

The theory of translocations, fusions, and fissions [12, 18] relies on the properties of sorting by reversals, thus most sorting reversals are preserving. Finally, transpositions are a more delicate matter since sorting transpositions are not (yet) characterized. Nevertheless, it is known that transpositions that increase the number of cycles – a desirable property when sorting permutations – have all their breakpoints in the same cycle [3]. Thus we have:

**Corollary 5.** *All transpositions that create two adjacencies are preserving.*

## 7 Conclusion

We have introduced a new similarity measure for permutations, based on the concept of conserved intervals. Conserved intervals have very interesting properties with respect to preserving the usual genome rearrangement operations. We believe that conserved intervals are a fundamental concept of rearrangement theory: they provide the unifying grounds to understand the variety of operations that are used to model genome evolution. Supported by recent results on the expected size of rearranged genome segments, one could go as far and claim that any rearrangement scenario that breaks conserved intervals is mathematical rambling without connection to evolutionary reality.

## References

1. Y. Ajana, J.-F. Lefebvre, E. R. M. Tillier, and N. El-Mabrouk. Exploring the set of all minimal sequences of reversals – an application to test the replication-directed reversal hypothesis. In *Proc. WABI 2002*, volume 2452 of *LNCS*, pages 300–315. Springer Verlag, 2002.

2. D. A. Bader, B. M. E. Moret, and M. Yan. A linear-time algorithm for computing inversion distance between signed permutations with an experimental study. *J. Comp. Biol.*, 8(5):483–492, 2001.
3. V. Bafna and P. A. Pevzner. Sorting by transpositions. *SIAM J. Disc. Math.*, 11(2):224–240, 1998.
4. A. Bergeron, S. Heber, and J. Stoye. Common intervals and sorting by reversals: A marriage of necessity. *Bioinformatics*, 18(Suppl. 2):S54–S63, 2002. (Proc. ECCB 2002).
5. A. Bergeron and J. Stoye. On the similarity of sets of permutations and its application to genome comparison. Report 2003-01, Technische Fakultät der Universität Bielefeld, 2003. (Available at `www.techfak.uni-bielefeld.de/~stoye/rpublications/report2003-01.pdf`).
6. M. Blanchette, T. Kunisawa, and D. Sankoff. Gene order breakpoint evidence in animal mitochondrial phylogeny. *J. Mol. Evol.*, 49(2):193–203, 1999.
7. J. L. Boore. Mitochondrial gene arrangement source guide. `www.jgi.doe.gov/programs/comparative/Mito_top_level.html`.
8. K. S. Booth and G. S. Lueker. Testing for the consecutive ones property, interval graphs and graph planarity using *PQ*-tree algorithms. *J. Comput. Syst. Sci.*, 13(3):335–379, 1976.
9. G. Bourque and P. A. Pevzner. Genome-scale evolution: Reconstructing gene orders in the ancestral species. *Genome Res.*, 12(1):26–36, 2002.
10. D. A. Christie. *Genome Rearrangement Problems.* PhD thesis, The University of Glasgow, 1998.
11. S. Hannenhalli and P. A. Pevzner. Transforming men into mice (polynomial algorithm for genomic distance problem). In *Proc. FOCS 1995*, pages 581–592. IEEE Press, 1995.
12. S. Hannenhalli and P. A. Pevzner. Transforming cabbage into turnip: Polynomial algorithm for sorting signed permutations by reversals. *J. ACM*, 46(1):1–27, 1999.
13. S. Heber and J. Stoye. Finding all common intervals of $k$ permutations. In *Proc. CPM 2001*, volume 2089 of *LNCS*, pages 207–218. Springer Verlag, 2001.
14. H. Kaplan, R. Shamir, and R. E. Tarjan. A faster and simpler algorithm for sorting signed permutations by reversals. *SIAM J. Computing*, 29(3):880–892, 1999.
15. J. D. Kececioglu and D. Sankoff. Efficient bounds for oriented chromosome inversion distance. In *Proc. CPM 1994*, volume 807 of *LNCS*, pages 307–325. Springer Verlag, 1994.
16. B. Larget, J. Kadane, and D. Simon. A Markov chain Monte Carlo approach to reconstructing ancestral genome rearrangements. Technical report, Carnegie Mellon University, Pittsburgh, 2002.
17. B. M. E. Moret, A. C. Siepel, J. Tang, and T. Liu. Inversion medians outperform breakpoint medians in phylogeny reconstruction from gene-order data. In *Proc. WABI 2002*, volume 2452 of *LNCS*, pages 521–536. Springer Verlag, 2002.
18. M. Ozery-Flato and R. Shamir. Two notes on genome rearrangements. *J. Bioinf. Comput. Biol.*, to appear.
19. D. Sankoff. Short inversions and conserved gene clusters. *Bioinformatics*, 18(10):1305–1308, 2002.
20. A. Siepel. An algorithm to find all sorting reversals. In *Proc. RECOMB 2002*, pages 281–290. ACM Press, 2002.
21. G. Tesler. Efficient algorithms for multichromosomal genome rearrangement. *J. Comput. Syst. Sci.*, 65(3):587–609, 2002.
22. T. Uno and M. Yagiura. Fast algorithms to enumerate all common intervals of two permutations. *Algorithmica*, 26(2):290–309, 2000.